ECONOMETRIC ANALYSIS

Bianca INOCENCIO
Glebs SOLOVJOVS
Rhita SQALLI

Possible models

1. **Fixed Effects (Within):** incorporates entity-specific variables into the model to account for unobserved variability across entities. It accounts for unobserved, time-invariant factors that could have an impact on the dependent variable. The FE estimator captures the within-entity variation over time by incorporating entity-fixed effects.
2. **Random Effects (GLS/FGLS):** takes into account both variance within and between entities. The entity-specific effects are assumed by the RE model to be uncorrelated with the independent variables, while they might be correlated with the error term. The correlation between the entity-specific effects and the error term is addressed using the GLS or FGLS estimator.
3. **Pooled OLS:** ignores entity-specific effects and time dependencies in favor of treating the panel data as a single cross-sectional dataset. Without making a distinction between different things or times, it calculates a single regression equation. Pooled OLS makes the assumption that there is no hidden heterogeneity or correlation between objects or across time.
4. **First Difference:** to address unobserved time-invariant heterogeneity in panel data analysis. It involves taking the difference between consecutive observations for each entity, effectively eliminating the time-invariant individual-specific factors from the analysis.

First, we set up our data panel with *xtset ivar tvar*, and then we look at the summary, we see that we have data for all individuals for all periods of time and so we can say that we have a **balanced panel**.

```
. xtset ivar tvar

Panel variable: ivar (strongly balanced)
 Time variable: tvar, 1 to 5
         Delta: 1 unit
```

```
. xtdescribe

  ivar:  1, 2, ..., 604                                 n =        604
  tvar:  1, 2, ..., 5                                   T =          5
         Delta(tvar) = 1 unit
         Span(tvar)  = 5 periods
         (ivar*tvar uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         5       5       5         5         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      604    100.00  100.00 |  11111
 ---------------------------+---------
      604    100.00         |  XXXXX
```

FE Estimator

The three R-sq measure the degree to which the LSDV estimated model can explain the within, the between, and the overall variation of the response variable y.

```
. xtreg y k l ,fe

Fixed-effects (within) regression               Number of obs     =      3,020
Group variable: ivar                            Number of groups  =        604

R-squared:                                      Obs per group:
     Within  = 0.1802                                         min =          5
     Between = 0.8810                                         avg =        5.0
     Overall = 0.5965                                         max =          5

                                                F(2, 2414)        =     265.24
corr(u_i, Xb) = 0.6866                          Prob > F          =     0.0000

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           k |   .4876862   .0235142    20.74   0.000      .441576    .5337964
           l |   .3895744   .0390815     9.97   0.000     .3129376    .4662111
       _cons |   1.533569   .0559373    27.42   0.000     1.423879    1.643259
-------------+----------------------------------------------------------------
     sigma_u |  .54061562
     sigma_e |  .56805995
         rho |    .475261   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(603, 2414) = 1.73                   Prob > F = 0.0000
```

We can see that the F-test yielded **Prob > F = 0.000** which means that the model is **significant and it has explanatory power**.

- The t-ratio test on parameters individually yielded P>|t| = 0.000 which means that coefficients βk and βl are different from 0 (both positive).
- All of the coefficients are individually significant and have the right sign..

sigma_e is the square root of the unbiased estimator for σLSDV.

_cons is the arithmetic mean for all the σ_{LSDV}. It is sizable and different from zero, however, we do not know how these components differ across individuals.

The test at the end tells us whether we can use POLS or not and disregard the panel structure altogether. The null hypothesis is α1=α2=α3=...=αn and in our case there is LH, so we cannot use **POLS**. We then check for the correlation of error u with y, which is positive.

```
. corr u y
(obs=3,020)

             |        u        y
-------------+------------------
           u |   1.0000
           y |   0.8128   1.0000
```

Robust FE and Two-way FE Estimator

Robust Variance:

We can see that almost 60% of the variation is described by the model, while standard error is comparable to the regular FE. All the coefficients are significant.

```
. xtreg y k l ,fe vce(robust)

Fixed-effects (within) regression               Number of obs     =      3,020
Group variable: ivar                            Number of groups  =        604

R-squared:                                      Obs per group:
     Within  = 0.1802                                         min =          5
     Between = 0.8810                                         avg =        5.0
     Overall = 0.5965                                         max =          5

                                                F(2, 603)         =     254.23
corr(u_i, Xb) = 0.6866                          Prob > F          =     0.0000

                                  (Std. err. adjusted for 604 clusters in ivar)
------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           k |   .4876862   .0246763    19.76   0.000     .4392243    .5361481
           l |   .3895744   .0377622    10.32   0.000      .315413    .4637357
       _cons |   1.533569   .0531233    28.87   0.000      1.42924    1.637898
-------------+----------------------------------------------------------------
     sigma_u |  .54061562
     sigma_e |  .56805995
         rho |    .475261   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Two-way estimator:

```
. quietly tabulate tvar,gen(time_d)

. xtreg y k l time_d1-time_d5 ,fe
note: time_d5 omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =      3,020
Group variable: ivar                            Number of groups  =        604

R-squared:                                      Obs per group:
     Within  = 0.3619                                         min =          5
     Between = 0.8823                                         avg =        5.0
     Overall = 0.6203                                         max =          5

                                                F(6, 2410)        =     227.76
corr(u_i, Xb) = 0.5951                          Prob > F          =     0.0000

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           k |   .4838285   .0208046    23.26   0.000     .4430317    .5246252
           l |   .3954738   .0345223    11.46   0.000     .3277775    .4631702
     time_d1 |   -.128353   .0289162    -4.44   0.000    -.1850561   -.0716499
     time_d2 |  -.2735202   .0288738    -9.47   0.000    -.3301403   -.2169001
     time_d3 |   .2459541   .0288943     8.51   0.000     .1892939    .3026144
     time_d4 |   .3795736   .0288825    13.14   0.000     .3229366    .4362107
     time_d5 |          0  (omitted)
       _cons |   1.483542   .0524324    28.29   0.000     1.380725    1.586359
-------------+----------------------------------------------------------------
     sigma_u |  .53877886
     sigma_e |   .5015906
         rho |  .53569962   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(603, 2410) = 2.21                   Prob > F = 0.0000

. testparm time_d1-time_d5

 ( 1)  time_d1 = 0
 ( 2)  time_d2 = 0
 ( 3)  time_d3 = 0
 ( 4)  time_d4 = 0

       F(  4,  2410) =  171.55
            Prob > F =    0.0000
```

When using the two-way FE the model explains the variation slightly better. From *testparm* we see that there is no homogeneity as the coefficients are different from each other. The model overall and all coefficients are significant.

RE Estimator

The coefficients on l and k are individually significant (P>t = 0.000 for both), collectively distinct from zero (Wald chi2(2)= 5149.06), and positive (0.6023521, 1.067278). Then, based on this output, l and k have a highly favorable effect on y. The assumption that corr(u_i, Xb) = 0—strict exogeneity of the regressors with respect to individual effects—underlies the random effect model.

```
. xtreg y k l ,re

Random-effects GLS regression                   Number of obs     =      3,020
Group variable: ivar                            Number of groups  =        604

R-squared:                                      Obs per group:
     Within  = 0.1568                                         min =          5
     Between = 0.9046                                         avg =        5.0
     Overall = 0.6387                                         max =          5

                                                Wald chi2(2)      =    5149.06
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
           k |   .6023521   .0217996    27.63   0.000     .5596257    .6450785
           l |   1.067278   .0207724    51.38   0.000     1.026565    1.107991
       _cons |   .5429267   .0287017    18.92   0.000     .4866725     .599181
-------------+----------------------------------------------------------------
     sigma_u |  .05502214
     sigma_e |  .56805995
         rho |  .00929461   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

- Now we test if there are random effects in the model.

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        y[ivar,t] = Xb + u[ivar] + e[ivar,t]

        Estimated results:
                         |       Var     SD = sqrt(Var)
                ---------+-----------------------------
                       y |   1.021903       1.010892
                       e |   .3226921       .5680599
                       u |   .0030274       .0550221

        Test: Var(u) = 0
                             chibar2(01) =     0.28
                          Prob > chibar2 =   0.2982
```

↳We do not reject the null hypothesis with Prob > chibar2 = 0.2982, we therefore better use the FE model.

```
. regress y l k ,vce(cluster ivar)

Linear regression                               Number of obs     =      3,020
                                                F(2, 603)         =    2374.54
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6387
                                                Root MSE          =     .60786

                                (Std. err. adjusted for 604 clusters in ivar)
------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           l |   1.073524    .021853    49.12   0.000     1.030607    1.116441
           k |   .6033576   .0224706    26.85   0.000     .5592274    .6474878
       _cons |   .5338313    .030894    17.28   0.000     .4731584    .5945042
------------------------------------------------------------------------------
```

Two-way RE

```
. xtreg y k l ib1.tvar

Random-effects GLS regression                   Number of obs     =      3,020
Group variable: ivar                            Number of groups  =        604

R-squared:                                      Obs per group:
     Within  = 0.3187                                         min =          5
     Between = 0.9045                                         avg =        5.0
     Overall = 0.6949                                         max =          5

                                                Wald chi2(6)      =    5676.26
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000
```

y	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
k	.5928659	.0199839	29.67	0.000	.5536982	.6320337
l	1.033179	.0200997	51.40	0.000	.9937842	1.072573
tvar						
2	-.151342	.0311327	-4.86	0.000	-.212361	-.0903231
3	.3655029	.0311182	11.75	0.000	.3045123	.4264934
4	.5104561	.0311199	16.40	0.000	.4494622	.5714501
5	.1313356	.03116	4.21	0.000	.0702632	.192408
_cons	.4241277	.034518	12.29	0.000	.3564738	.4917817
sigma_u	.13132871					
sigma_e	.5015906					
rho	.06415417	(fraction of variance due to u_i)				

The coefficients on l and k are individually significant (P>t = 0.000 for both), collectively different from zero (Wald chi2(2)= 5676.26), and positive (.5928659, 1.033179). We also notice the three R-squared measurements, and since none of them were obtained using OLS regression, we can only interpret them here as squared correlation coefficients.

- **Within = 0.3187** squared correlation of 31.87% between y and fitted values both in individual-mean deviations
- **Between = 0.9045** squared correlation of 90.45% for both group means of y and the fitted values
- **Overall = 0.6949** squared correlation between the untransformed values of y and the fitted values, which is 69.49%.

This estimator performs similarly between than the between estimator, similarly better within than the within fixed-effects estimator, and slightly better overall, according to the R-squared.

FD estimator: Autocorrelation and heteroskedasticity

- Serial correlation:

```
. xtserial y l k

Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,      603) =     35.974
           Prob > F =      0.0000
```

↳ We can observe there is first-order autocorrelation.

- Test for heteroskedasticity in the FE setting:

```
. quietly xtreg y l k ,fe

. xttest3

Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (604)  =   26871.23
Prob>chi2 =      0.0000
```

↳Strong evidence against the null hypothesis of homoskedasticity, indicating the presence of groupwise heteroskedasticity. We need to address the issue by using the robust variance.

- We then look at the FD estimator:

```
. xtserial y l k ,output

Linear regression                               Number of obs     =      2,416
                                                F(2, 603)         =     205.87
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1952
                                                Root MSE          =     .77132

                              (Std. err. adjusted for 604 clusters in ivar)
```

D.y	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
l						
D1.	.3624337	.042283	8.57	0.000	.2793938	.4454735
k						
D1.	.5057574	.0269717	18.75	0.000	.4527874	.5587273

```
Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,      603) =     35.974
           Prob > F =      0.0000
```

↳Once again, we see that there is first-order correlation.

Hausman Test

● We apply the Hausman test to compare the Fixed Effects and the Random Effects models:

↳ We can observe that the null hypothesis is rejected, meaning that the random effects (RE) estimator and the fixed effects (FE) estimator are consistent. We can then infer that the Fixed Effect model is better suited for our study than the Random Effect.

```
. quietly xtreg y l k ,fe

. estimates store fe

. quietly xtreg y l k ,re

. estimates store re

. hausman fe re ,sigmaless

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fe           re          Difference       Std. err.
-------------+----------------------------------------------------------------
           l |   .3895744     1.067278       -.6777034        .0338676
           k |   .4876862     .6023521       -.1146659        .0115769
------------------------------------------------------------------------------
                          b = Consistent under H0 and Ha; obtained from xtreg.
           B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

    chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            = 407.69
Prob > chi2 = 0.0000
```

Robust Hausman Test

- First, we do the regression with group means.

```
. foreach x of varlist l k {
  2. bysort ivar: egen gm`x'=mean(`x')
  3. }

. xtreg y l k gml gmk, re vce(cluster ivar)

Random-effects GLS regression                   Number of obs     =      3,020
Group variable: ivar                            Number of groups  =        604

R-squared:                                      Obs per group:
     Within  = 0.1802                                         min =          5
     Between = 0.9049                                         avg =        5.0
     Overall = 0.6817                                         max =          5

                                                Wald chi2(4)      =    6285.55
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000
```

(Std. err. adjusted for 604 clusters in ivar)

y	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
l	.3895744	.0377747	10.31	0.000	.3155373	.4636114
k	.4876862	.0246845	19.76	0.000	.4393055	.5360668
gml	.8534943	.0487775	17.50	0.000	.7578921	.9490965
gmk	.1374399	.0509745	2.70	0.007	.0375316	.2373482
_cons	.290738	.0310231	9.37	0.000	.2299338	.3515422
sigma_u	.05502214					
sigma_e	.56805995					
rho	.00929461	(fraction of variance due to u_i)				

- Now we do the test:

```
. testparm gmk gml

 ( 1)  gml = 0
 ( 2)  gmk = 0

           chi2(  2) =  415.36
         Prob > chi2 =    0.0000
```

↳ The Robust Hausman test leads us to the same conclusion than the standard Hausman test: the p-value is 0, we can reject the null hypothesis that the RE estimator and the FE estimator are consistent and choose the FE model, even after accounting for heteroskedasticity and/or serial correlation.

Conclusion

Comparing the four specifications (Fixed effects, Random effects, Pooled OLS and First Difference) we can conclude that the Fixed Effects with robust variance is the best model to estimate the production function with panel data on y = output, x = capital; labor.

We eliminate the POLS as from the output of the FE regression we see that there is heterogeneity and so we cannot get rid of the panel structure.

Looking at the Wooldridge test for first-order autocorrelation we can see that it is present as H_0 is rejected. Thus, we look at the robust Hausman test that rejects the H_0 indicating that FE is more effective than RE in this situation. This is also supported by the test for random effects which indicates that they are absent from the model.

We eliminate the FD model as it requires the heterogeneity to be time-invariant, however, in our case it varies with time. This can be understood from the Hausman test that indicates that FE is better than RE and thus the heterogeneity varies with time.

As there is autocorrelation we also eliminate two-way FE as it assumes no correlation between error terms.

Lastly, to address the issue of heteroskedasticity and autocorrelation we opt for the robust variance version of FE.

The model shows us that there is (obviously) a positive relationship between labor and output, but that this relationship is less strong than the one between capital and output, which is also positive. In addition to that, we can say that the most accurate specification is the **Fixed Effects (Within) with robust variance**.